

EXEMPLAR-BASED LINEAR DISCRIMINANT ANALYSIS FOR ROBUST OBJECT TRACKING

Changxin Gao, Feifei Chen, Jin-Gang Yu, Rui Huang, Nong Sang

Science and Technology on Multi-spectral Information Processing Laboratory,
School of Automation, Huazhong University of Science and Technology, Wuhan, 430074, China
cgao@hust.edu.cn

ABSTRACT

Tracking-by-detection has become an attractive tracking technique, which treats tracking as a category detection problem. However, the task in tracking is to search for a specific object, rather than an object category as in detection. In this paper, we propose a novel tracking framework based on exemplar detector rather than category detector. The proposed tracker is an ensemble of exemplar-based linear discriminant analysis (ELDA) detectors. Each detector is quite specific and discriminative, because it is trained by a single object instance and massive negatives. To improve its adaptivity, we update both object and background models. Experimental results on several challenging video sequences demonstrate the effectiveness and robustness of our tracking algorithm.

Index Terms— Exemplar, Linear Discriminant Analysis (LDA), Object tracking, Model updating

1. INTRODUCTION

Visual tracking plays a key role in many computer vision applications, such as surveillance, HCI, video editing, *etc.* It has been studied intensively during the past decades [1], the problem in general still remains challenging due to various factors such as appearance, pose, and scale change of objects, occlusion of objects, illumination variations, cluttered scenes, presence of similar objects, *etc.*

Recently, tracking-by-detection method has become an attractive tracking technique [2, 3, 4, 5, 6], which treats tracking as a classification problem and trains a detector to separate the object from the background. Good performance has been shown following this strategy, by borrowing some techniques from object detection methods [3, 5, 7, 8, 9, 10]. Furthermore, Stalder *et al.* discussed the relationship of tracking, detection, and recognition in [11]. However, the task in tracking is different from that in detection, that is, finding a specific object instance in tracking, while finding an object category in detection. Therefore, we suppose that tracking should be based on object exemplar rather than category. That is to say, we should design a specific, and

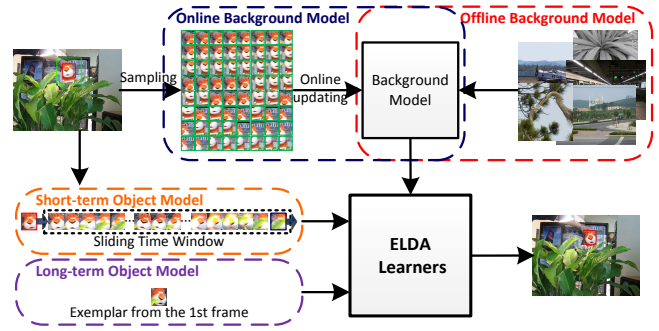


Fig. 1. An overview of the ELDA tracking algorithm. ELDA tracker consists of two object models (long-term object model and short-term object model), and two background models (off-line background model and online background model). The proposed method builds a single ELDA learner for each sample in object model, with both off-line and online background samples. The figure is best viewed in color.

more restrictive detector for the object instance to be tracked, rather than a category-based detector.

To this end, we present an Exemplar-based Linear Discriminant Analysis (ELDA) model for visual tracking. Exemplar-based learner is supposed to be extremely discriminative, because it is trained by a single object instance and massive amounts of negative samples. Malisiewicz *et al.* proposed exemplar-based support vector machine (SVM) algorithm for object detection in [12]. However, training an exemplar-based SVM, by mining for hard negative exemplars, is quite expensive. Linear Discriminant Analysis (LDA) technique is introduced to improve the speed of training and testing [13], which enables exemplar-based method to be used in tracking task.

ELDA algorithm consists of two parts, object model and background model, as in Fig. 1. To achieve good adaptivity of the proposed method, we update both of them. The object model should give full play to the role of each object exemplar during the tracking process, to handle with the variety of the object. Although training an ELDA detector is very cheap, taking all the exemplars to build the object

model is infeasible, especially in a very long-time tracking. We only use the exemplars in a predefined sliding time window in this work. On the other hand, the first frame is very important in tracking, because it includes the precise labels. Thus, we use this exemplar during the whole tracking process, called long-term object model. Similarly, we call the previous one short-term object model.

To train a discriminative ELDA detector, vast amounts of negative samples are required. However, we know that, it is difficult and time-consuming to obtain in the procedure of tracking. Thus, we build an initial background model by collecting a sufficient large scale negative set from some natural images with an off-line manner. On the other hand, the background information just around the object instance is critical for tracking in both discrimination and adaptivity. Accordingly, we also update background model by an online manner besides off-line one.

To sum up, we present a novel visual tracking framework called ELDA, which is quite discriminative due to exemplar-based learner training by a single object instance and massive negatives, and is quite adaptivity due to online updating both object and background models, as shown in Fig. 1. We apply this algorithm to visual tracking on several public video sequences and find the results quite promising.

2. RELATED WORK

Prior approaches to object tracking can be roughly divided into two broad categories for build tracking model, namely, generative and discriminative. We first point the readers to a survey work [14] and a recent benchmark work [1]. Our method is a discriminative tracking model, we therefore present some other previous work on this topic [15, 16, 17, 18, 19, 20, 21]. As shown in these works, most discriminative tracking methods are based on haar-like feature representation and online boosting classifier. However, in object detection, the most popular framework is based on HOG feature [22] and linear SVM or LDA. Furthermore, Struck [6], as one of the most comparative tracker, shows the advantages of SVM classifier in tracking. Motivated by these works, we introduce this detection framework into tracking problem. However, the most difference of our method with the state-of-the-art tracking-by-detection methods is that the detectors in our method are trained by an exemplar-based classifiers, rather than category-based classifiers.

Exemplar-SVM detection methods have recently become particularly popular, due to its discrimination ability. Exemplar-based SVM, first introduced in [12], shows good performance by learning an object model from each single object exemplar. However, training an exemplar-SVM, by mining for hard negative exemplars, is quite expensive. To resolve this problem, Ref. [13] applies Linear Discriminant Analysis (LDA) technique to speed up. This framework has

been widely used in many computer vision applications, *i.e.*, object detection [12, 23], image retrieval [24], mid-level representation discovery for scene classification [25], and action classification [26], *etc.* Note that, HOG and exemplar-SVM (or LDA) framework makes effectiveness of these methods. In view of this, we extend exemplar-based LDA method to online tracking case, by updating both object and background models.

3. ELDA TRACKING METHOD

In this section, we introduce the ELDA tracking algorithm, and focus on the process of building ELDA detector, and updating both object model and background model. Fig. 1 is an overview of the proposed approach.

3.1. ELDA detector

First we present the typical tracking-by-detection algorithms, which train a detector to distinguish a target object from its local background. Specifically, given a bounding box position c_k (initial position or tracked result position) in frame k , a tracker first labels the samples in a predefined training area R_t , with the size of radius r , to $y = 1$ and $y = -1$ for positive and negative samples respectively; then trains a classifier using the feature representations of the samples X and corresponding labels y ; final tracking result is combined by classifying the samples in detection area R_d in frame $k + 1$.

In ELDA tracking algorithm, we only take the sample exactly at position c_k as positive sample, rather than the samples in a very close area around c_k as in typical tracking-by-detection method. The representation of the positive sample in frame k is denoted as X_k^p . Then we train an LDA classifier for X_k^p , using the covariance matrix Σ_k and means μ_k^n of a negative dataset, which will be introduced in Sec. 3.3. The final ELDA classifier can be written as:

$$H_k(X) = \text{sign}(\omega_k^T \times X + b_k) \quad (1)$$

where ω_0 is the threshold, the weights can be calculate by:

$$\omega_k = \Sigma^{-1}(X_k^p - \mu_k^n) \quad (2)$$

3.2. Object Model

We build object model for each positive exemplar, thus, the key is how to choose the positive samples. The first frame with the precise labels is critical in tracking. Thus, we use the object exemplar in the first frame during the whole tracking process, called long-term object model, denoted as $H_1(X)$. On the other hand, to improve the adaptivity to the variety of object appearance, we would better build object model by applying as many as possible ELDA detectors in theory. However, it's not a good solution in practical

application, especially for long-time tracking. In this paper, we simply set a time window TM to choose positives. Only the samples in past TM frames from frame k are used to build object model, namely, X_i^p , $i \in [l, k]$, $l = \max(2, k - TM + 1)$, and we call it short-term object model. The weights of the ELDA detectors $H_k(X)$ are determined by a semi-supervised way using the long-term detector $H_1(X)$ as the prior, then set the weight λ_k to $H_k(X)$ as follows:

$$\lambda_k = \frac{H_1(X_k^p)}{H_1(X_1^p)} \quad (3)$$

Accordingly, the object model can be defined as:

$$M_k^O(X) = \lambda_1 H_1(X) + \sum_i \lambda_i H_i(X) \quad (4)$$

3.3. Background Model

In exemplar-based framework, the background model M_k^B can be denoted as (Σ_k, μ_k^n) according to Eq. 1 and Eq. 2. In the tracking case, the negative samples in a ring area centered at the object position are critical, where we sample online negatives as most of other tracking-by-detection approaches [7, 4]. On the other hand, the huge number of negatives are the guarantee of discrimination ability of ELDA. However, it is difficult and time-consuming to obtain lots of negatives in the procedure of tracking. Therefore, we utilize a strategy to build a background model with large scale negatives, collecting by both off-line and online manner. To build the off-line model, we first collect massive amounts of negative samples from some natural images, and then calculate the background model (Σ_0, μ_0) as initial model $M_0^B = (\Sigma_0, \mu_0)$, X_0 is the representations of all negative samples.

Online background model is used to improve the adaptivity by some negative samples quite relevant to the tracking task, that is in the ring area mentioned above. In frame k , we calculate online model M_k^{Bon} using the negative samples X_k^n . The final background model in frame k is incrementally calculated using M_{k-1}^B and M_k^{Bon} according to definition of covariance matrix and means.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

In this section, we evaluate our ELDA tracker on ten public available benchmark video sequences used in previous works [4, 6, 1], namely, david indoor (david), sylvester, singer2, coke can (coke), girl, david outdoor (david3), suv, liquor, woman, and tiger1. These videos are very difficult to track, because of various challenges, such as, occlusion of objects, illumination variation, appearance change of objects, rotation and scale change of objects, clutter scenes, presence

of similar objects, *etc.* Some samples corresponding to these challenges can be seen in Fig. 2, and more details are introduced in [1]. All the settings of videos are same as in [1], *e.g.*, tiger1 starts from frame 6.

In our experiments, ELDA tracker is compared with six state-of-the-art tracking-by-detection algorithms, the fragment tracker (Frag) [27], the online boosting tracker (OAB) [7], the visual tracking decomposition algorithm (VTD) [17], the multiple instance learning tracker (MIL) [4], the incremental visual tracking method (IVT) [28], and Struck [6].

4.2. Implementation details

For the representation, we use HOG feature (8×8 cells with 9 orientations) in this work. Thus, the resulting feature is $8 \times 8 \times 4 \times 9 = 2304$ dimensions. To build short-term object model, we set the size of time window $TM = 500frames$, which is determined experimentally. To build off-line background model, we collected more than 1000,000 patches (64×64 pixels) by randomly sampling on the 5096 images of PASCAL VOC 2008 dataset [29]. Then HOG feature is extracted to build initial background model M_0^B ; and the online negatives are sampled in the ring area with $5 < d \leq 30$. The detect area R_d is also set to 30.

4.3. Quantitative Comparison

Two common evaluation criteria are used for quantitative comparison, namely, center location error (CLE) and success rate (SR). First we define these two criteria briefly. For each frame, the result is denoted as tracked bounding box B_T and center location C_T , which of ground truth is B_G and C_G respectively. CLE is defined as the average Euclidean distance (in pixels) between C_T and C_G . SR is defined as the rate of successful frames in total frames. A tracked result is considered to be successful if the overlap ratio $\frac{area(B_T \cap B_G)}{area(B_T \cup B_G)}$ is larger than 0.5.

Table 1 and Table 2 report the comparison results of ELDA and other six state-of-the-art trackers in terms of average center location error and success rate. It can be seen that, ELDA tracker outperforms other trackers on 5 out of 10 videos, and obtains 8 best or second best scores out of 10 videos in terms of both average center location error and success rate. Most exiting, ELDA tracker, over all, performs well against other six state-of-the-art algorithms. Note that only one average center location error is bigger than 20 pixels in our results, which demonstrate the proposed method works robustly.

To highlight the superior performance of the ELDA tracker, we show some images with comparison tracked bounding box in Fig. 2 under lots of special challenges, *e.g.*, heavy occlusion, illumination variations, appearance changes, rotation and scale changes, background cluttering,

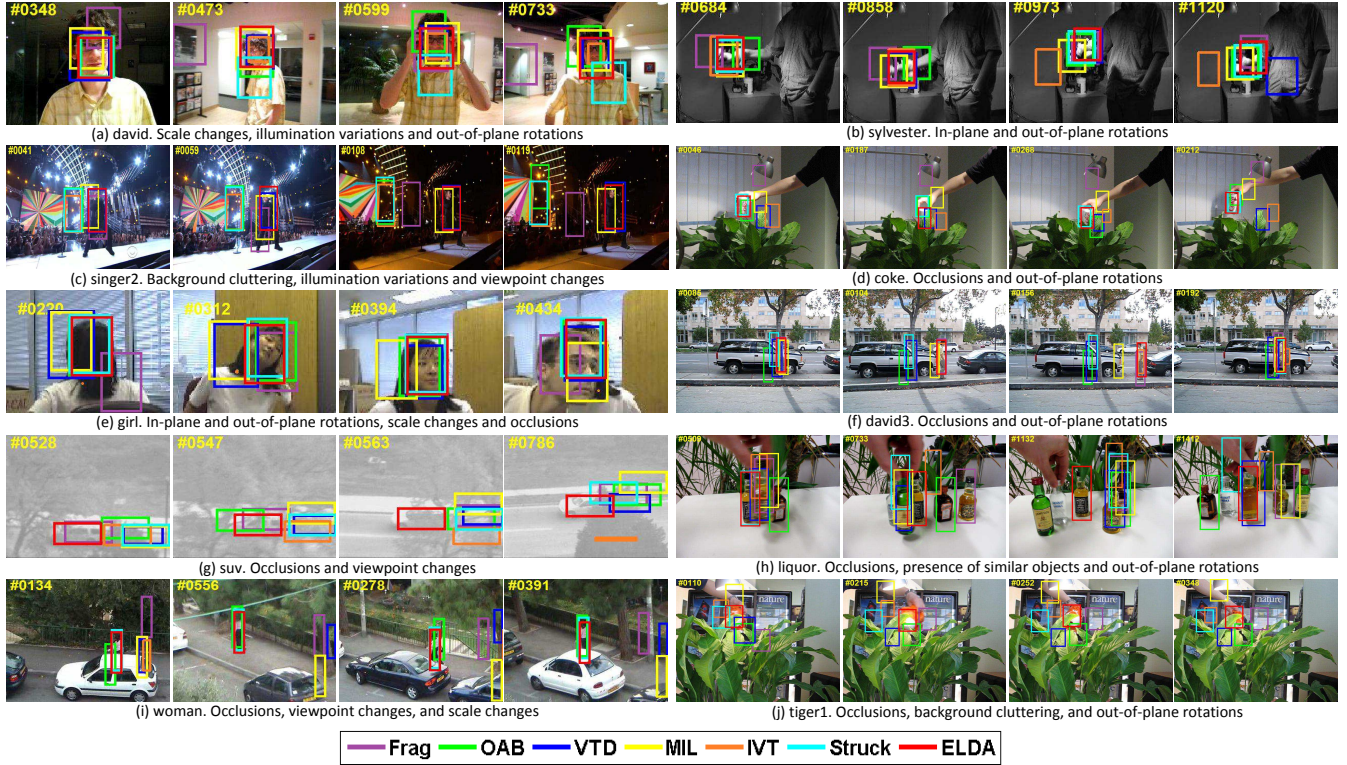


Fig. 2. Tracked bounding box results comparisons of 7 trackers in 10 videos under some special challenges.

Table 1. Average center location error (in pixels) comparison of the 7 trackers on 10 videos. Bold and underlined values indicates best and second best performance.

Sequence	Frag	OAB	VTD	MIL	IVT	Struck	ELDA
david	82.1	21.7	11.6	16.9	4.8	42.8	<u>7.9</u>
sylvester	15.0	<u>14.8</u>	19.6	15.2	34.2	6.3	17.3
singer2	88.6	<u>185.9</u>	43.7	<u>22.5</u>	175.5	174.3	9.3
coke	124.8	35.9	68.6	46.7	83.0	12.1	<u>14.3</u>
girl	20.7	<u>3.7</u>	8.6	13.7	22.5	2.6	3.7
david3	<u>13.6</u>	83.4	66.7	29.7	51.9	106.5	6.8
suv	42.0	<u>30.5</u>	57.2	82.2	57.3	49.8	9.7
liquor	99.6	68.6	<u>60.2</u>	141.9	118.5	91.0	33.2
woman	111.9	31.4	118.9	125.3	176.5	4.2	<u>5.6</u>
tiger1	<u>74.3</u>	94.9	107.3	108.9	106.6	128.4	16.9
MEAN	67.3	57.1	<u>56.2</u>	60.3	83.1	61.8	12.5

presence of similar objects occur. To present the tracking results frame by frame, we also give the corresponding tracking error of 7 trackers on 10 video sequences in Fig. 3. It shows the good performance of ELDA tracker in both accuracy and adaptivity.

5. CONCLUSIONS AND FUTURE WORK

The task in tracking is to search for a specific object instance, rather than an object category as in detection. In view of this, we proposed a new tracking framework based on exemplar detector rather than category detector. We build

Table 2. Success rate comparison of the 7 trackers on 10 videos. Bold and underlined values indicates best and second best performance.

Sequence	Frag	OAB	VTD	MIL	IVT	Struck	ELDA
david	0.12	0.15	<u>0.68</u>	0.23	0.79	0.24	0.61
sylvester	0.68	0.68	0.80	0.55	0.68	0.93	0.79
singer2	0.20	0.09	0.03	0.48	0.04	0.04	0.94
coke	0.03	0.11	0.17	0.12	0.13	0.94	<u>0.66</u>
girl	0.54	0.46	<u>0.94</u>	0.29	0.19	0.98	<u>0.94</u>
david3	<u>0.81</u>	0.34	0.48	0.68	0.63	0.34	0.99
suv	0.71	<u>0.76</u>	0.55	0.13	0.44	0.57	0.87
liquor	0.37	0.48	<u>0.58</u>	0.20	0.21	0.41	0.85
woman	0.18	<u>0.61</u>	0.18	0.19	0.18	0.93	0.93
tiger1	<u>0.31</u>	<u>0.09</u>	0.12	0.10	0.08	0.18	0.83
MEAN	0.40	0.42	0.46	0.30	0.34	<u>0.56</u>	0.84

ELDA tracker by both updating object and background models. Promising results on challenging video sequences demonstrate that our method outperforms the state-of-the-art tracking algorithms. We are considering the following for the future work. First, in our current tracker, updating strategy of object model with a predefined time widow is very simple. To further improve the adaptivity, we are looking for a more effective updating method. Second, due to the successful of part-based model in object detection [30], we will study part-based tracking approach, to deal with some challenges, e.g., the occlusion, deformation.

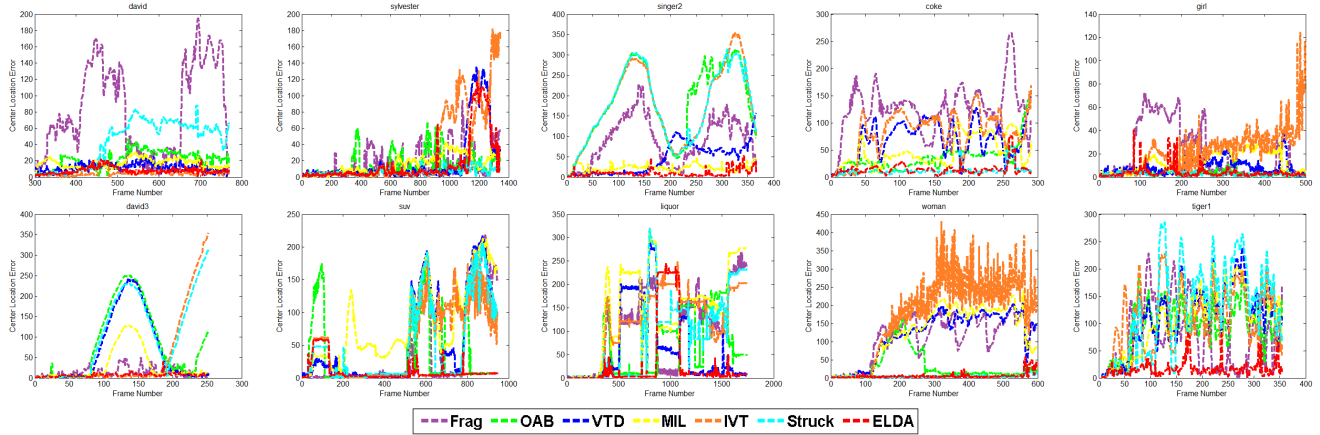


Fig. 3. Tracking error (in pixels) of 7 trackers on 10 video sequences. The figure is best viewed in color.

Acknowledgments

This work is supported by the Project of the National Natural Science Foundation of China No.61271328, and the Fundamental Research Funds for the Central Universities No. HUST2013TS132.

6. REFERENCES

- [1] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *CVPR*, 2013.
- [2] S. Avidan, "Support vector tracking," in *CVPR*, 2001.
- [3] S. Avidan, "Ensemble tracking," in *CVPR*, 2005.
- [4] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *CVPR*, 2009.
- [5] R.T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE T. PAMI*, vol. 27, no. 10, pp. 1631–1643, 2005.
- [6] S. Hare, A. Saffari, and P. H. S Torr, "Struck: Structured output tracking with kernels," in *ICCV*, 2011.
- [7] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *BMVC*, 2006.
- [8] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," *ECCV*, 2008.
- [9] B. Zeisl, C. Leistner, A. Saffari, and H. Bischof, "Online semi-supervised multiple-instance boosting," in *CVPR*, 2010.
- [10] C. Gao, N. Sang, and R. Huang, "Online transfer boosting for object tracking," in *ICPR*, 2012.
- [11] S. Stalder, H. Grabner, and L. Van Gool, "Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition," in *ICCV Workshops*, 2009.
- [12] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *ICCV*, 2011.
- [13] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative decorrelation for clustering and classification," in *ECCV*, 2012.
- [14] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm Computing Surveys*, vol. 38, no. 4, 2006.
- [15] Y. Wu and J. Fan, "Contextual flow," in *CVPR*, 2009.
- [16] Z. Kalal, J. Matas, and K. Mikołajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *CVPR*, 2010.
- [17] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *CVPR*, 2010.
- [18] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," in *CVPR*, 2010, pp. 723–730.
- [19] T. B. Dinh, N. Vo, and G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *CVPR*, 2011.
- [20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision—ECCV 2012*, 2012.
- [21] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *ECCV*, 2012.

- [22] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.
- [23] G. Shakhnarovich S. Maji, “Part discovery from partial correspondence,” in *CVPR*, 2013.
- [24] T. Malisiewicz, A. Shrivastava, A. Gupta, and A. A. Efros, “Exemplar-svms for visual object detection, label transfer and image retrieval,” in *ICML*, 2012.
- [25] M. Juneja, A. Vedaldi, C.V. Jawahar, and A. Zisserman, “Blocks that shout: Distinctive parts for scene classification,” in *CVPR*, 2013.
- [26] C. Doersch, A. Gupta, and A. A. Efros, “Mid-level visual element discovery as discriminative mode seeking,” in *NIPS*, 2013.
- [27] A. Adam, E. Rivlin, and I. Shimshoni, “Robust fragments-based tracking using the integral histogram,” in *CVPR*, 2006.
- [28] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, “Incremental visual tracking,” .
- [29] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results,” <http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html>.
- [30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.